```
library("readr")
library("tidyr")
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v dplyr   1.0.7
## v tibble  3.1.4     v stringr 1.4.0
## v purrr   0.3.4     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("dplyr")
library("ggplot2")
```

```
#Source of file:-
#kaggle.com/sidtwr/videogames-sales-dataset?select=Video_Games_Sales_as_at_22_Dec_2016.csv


#Variable Definition:

#video_games_sales_data stores the dataset as it is.
dataset<-read_csv("D:/Documents/IPL Matches 2008-2020.csv")
```

```
## Rows: 812 Columns: 17
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (14): city, date, player_of_match, venue, team1, team2, toss_winner, tos...
## dbl  (3): id, neutral_venue, result_margin
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
dataset
```

```
## # A tibble: 812 x 17
##        id city       date      player_of_match venue  neutral_venue team1  team2
##     <dbl> <chr>      <chr>     <chr>           <chr>          <dbl> <chr>  <chr>
##  1 335982 Bangalore  4/18/2008 BB McCullum     M Chi~             0 Royal~ Kolk~
##  2 335983 Chandigarh 4/19/2008 MEK Hussey      Punja~             0 Kings~ Chen~
##  3 335984 Delhi      4/19/2008 MF Maharoof     Feroz~             0 Delhi~ Raja~
##  4 335985 Mumbai     4/20/2008 MV Boucher      Wankh~             0 Mumba~ Roya~
##  5 335986 Kolkata    4/20/2008 DJ Hussey       Eden ~             0 Kolka~ Decc~
##  6 335987 Jaipur     4/21/2008 SR Watson       Sawai~             0 Rajas~ King~
##  7 335988 Hyderabad  4/22/2008 V Sehwag        Rajiv~             0 Decca~ Delh~
##  8 335989 Chennai    4/23/2008 ML Hayden       MA Ch~             0 Chenn~ Mumb~
##  9 335990 Hyderabad  4/24/2008 YK Pathan       Rajiv~             0 Decca~ Raja~
## 10 335991 Chandigarh 4/25/2008 KC Sangakkara   Punja~             0 Kings~ Mumb~
## # ... with 802 more rows, and 9 more variables: toss_winner <chr>,
## #   toss_decision <chr>, winner <chr>, result <chr>, result_margin <dbl>,
## #   eliminator <chr>, method <chr>, umpire1 <chr>, umpire2 <chr>
```

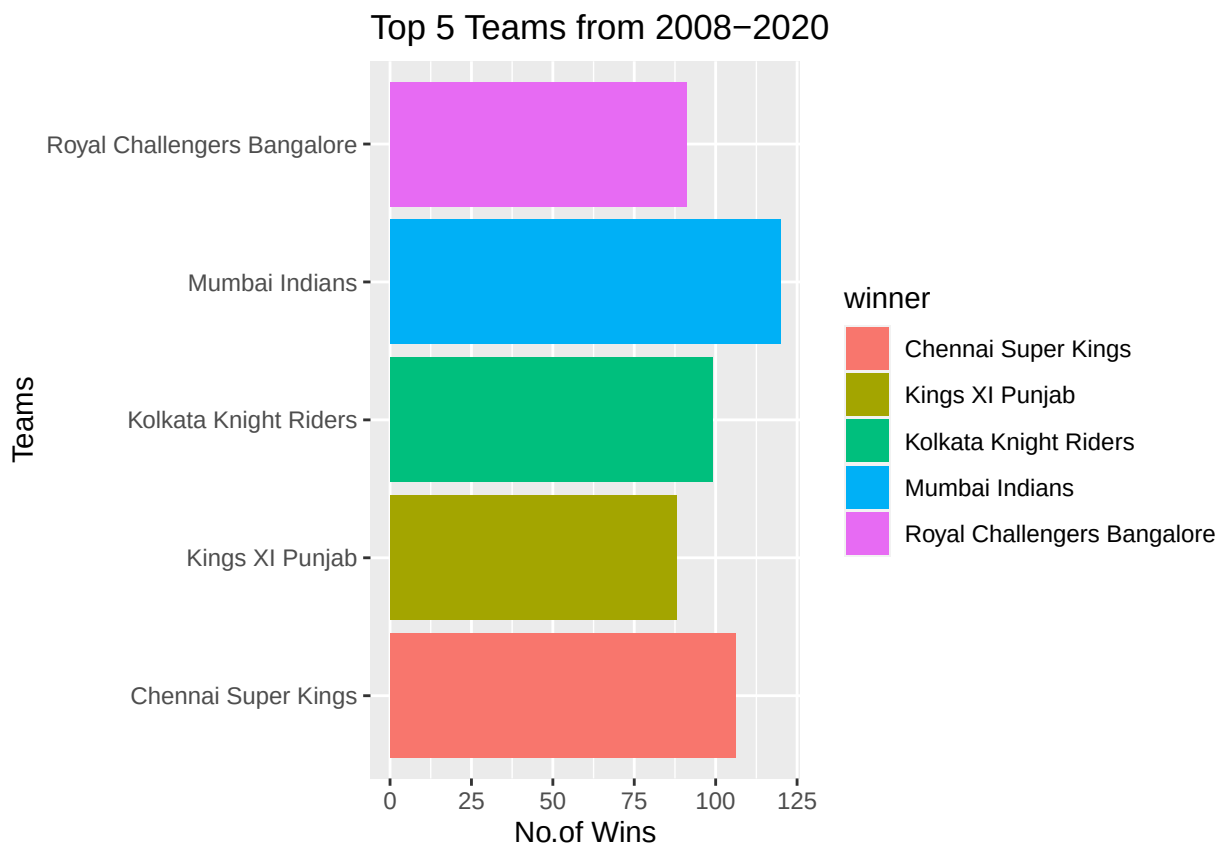
```
data_sorted<- dataset %>%
    group_by(winner) %>%
```

```
     summarise(Count = n()) %>%
      arrange(desc(Count))
head(data_sorted,5)
```

```
## # A tibble: 5 x 2
##   winner                       Count
##   <chr>                        <int>
## 1 Mumbai Indians                 120
## 2 Chennai Super Kings            106
## 3 Kolkata Knight Riders           99
## 4 Royal Challengers Bangalore     91
## 5 Kings XI Punjab                 88
```

```
data_sorted<-head(data_sorted,5)
```

```
ggplot(data = data_sorted,aes(y,x=Count)) +
  geom_bar(stat='identity',aes(y=winner,fill=winner)) +labs(x= "No.of Wins",y="Teams",
title="Top 5 Teams from 2008-2020")
```




```
dataset$date <- as.POSIXct(dataset$date,format = "%m/%d/%Y")
dataset$date <- format(dataset$date, format="%Y")
dataset
```

```
## # A tibble: 812 x 17
##        id city       date  player_of_match venue    neutral_venue team1   team2
##     <dbl> <chr>      <chr> <chr>           <chr>            <dbl> <chr>   <chr>
##  1 335982 Bangalore  2008  BB McCullum     M Chinn~             0 Royal ~ Kolka~
```

```
##  2 335983 Chandigarh 2008  MEK Hussey      Punjab ~             0 Kings ~ Chenn~
##  3 335984 Delhi      2008  MF Maharoof     Feroz S~             0 Delhi ~ Rajas~
##  4 335985 Mumbai     2008  MV Boucher      Wankhed~             0 Mumbai~ Royal~
##  5 335986 Kolkata    2008  DJ Hussey       Eden Ga~             0 Kolkat~ Decca~
##  6 335987 Jaipur     2008  SR Watson       Sawai M~             0 Rajast~ Kings~
##  7 335988 Hyderabad  2008  V Sehwag        Rajiv G~             0 Deccan~ Delhi~
##  8 335989 Chennai    2008  ML Hayden       MA Chid~             0 Chenna~ Mumba~
##  9 335990 Hyderabad  2008  YK Pathan       Rajiv G~             0 Deccan~ Rajas~
## 10 335991 Chandigarh 2008  KC Sangakkara   Punjab ~             0 Kings ~ Mumba~
## # ... with 802 more rows, and 9 more variables: toss_winner <chr>,
## #   toss_decision <chr>, winner <chr>, result <chr>, result_margin <dbl>,
## #   eliminator <chr>, method <chr>, umpire1 <chr>, umpire2 <chr>
```

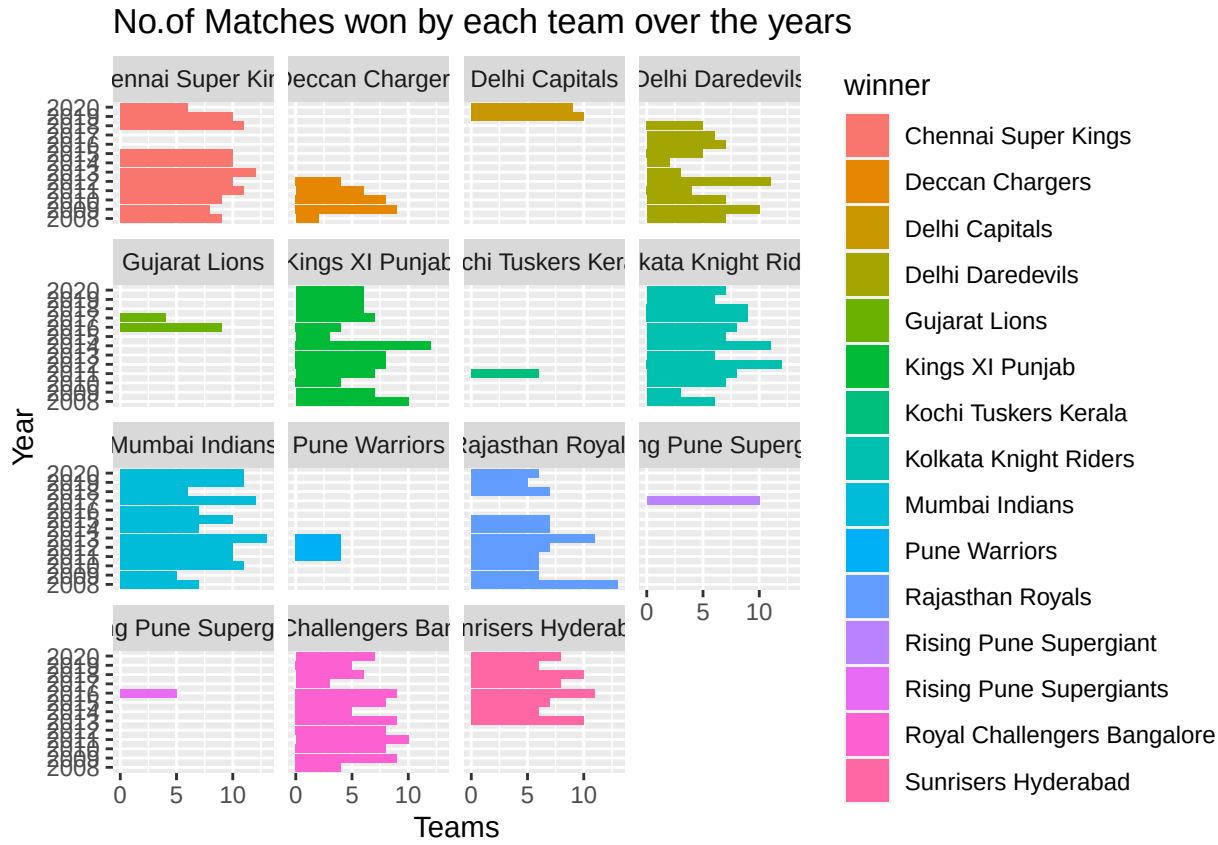
```
data_sorted<- dataset %>%
    group_by(date,winner)%>%
      summarise(Count = n()) %>%
      arrange(desc(Count))
```

```
## `summarise()` has grouped output by 'date'. You can override using the `.groups` argument.
```

```
data_sorted
```

```
## # A tibble: 108 x 3
## # Groups:   date [13]
##    date  winner                Count
##    <chr> <chr>                 <int>
##  1 2008  Rajasthan Royals          13
##  2 2013  Mumbai Indians            13
##  3 2012  Kolkata Knight Riders     12
##  4 2013  Chennai Super Kings       12
##  5 2014  Kings XI Punjab           12
##  6 2017  Mumbai Indians            12
##  7 2010  Mumbai Indians            11
##  8 2011  Chennai Super Kings       11
##  9 2012  Delhi Daredevils          11
## 10 2013  Rajasthan Royals          11
## # ... with 98 more rows
```

```
ggplot(data = data_sorted,aes(y,x=Count)) +
  geom_bar(stat='identity',aes(y=date,fill=winner)) +labs(x="Teams",y="Year",
title="No.of Matches won by each team over the years") + facet_wrap(~winner)
```



PART -B: Problems 3–5 use the PimaIndiansDiabetes2 dataset from the mlbench package. You do not need to partition the dataset for any of the problems in Part A.

Problem 3: We would like to know if there is difference in blood pressure between people with diabetes and people without diabetes. First remove missing values from the data using 'na.omit()'. Then fit a model for blood pressure using diabetes as the only explanatory variable. Perform model diagnostics to check for any violations of model assumptions. Visualize the relationship between blood pressure and diabetes. State the null and alternative hypotheses, choose an alpha value, and state the p-value and your conclusions.

```
library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.1.2
```

```
library(modelr)
```

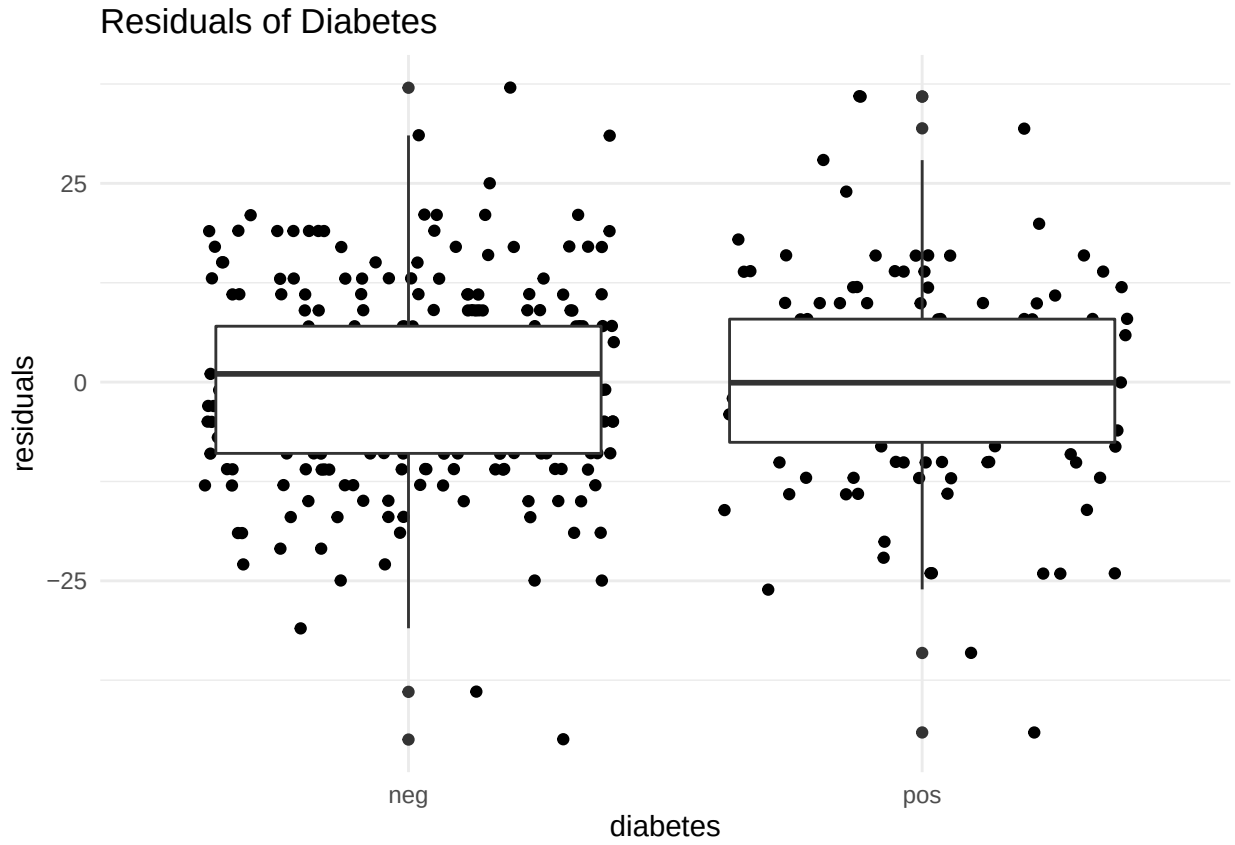
```
## Warning: package 'modelr' was built under R version 4.1.2
```

```
data(PimaIndiansDiabetes2)
model_data <- as_tibble(na.omit(PimaIndiansDiabetes2))
model_data
```

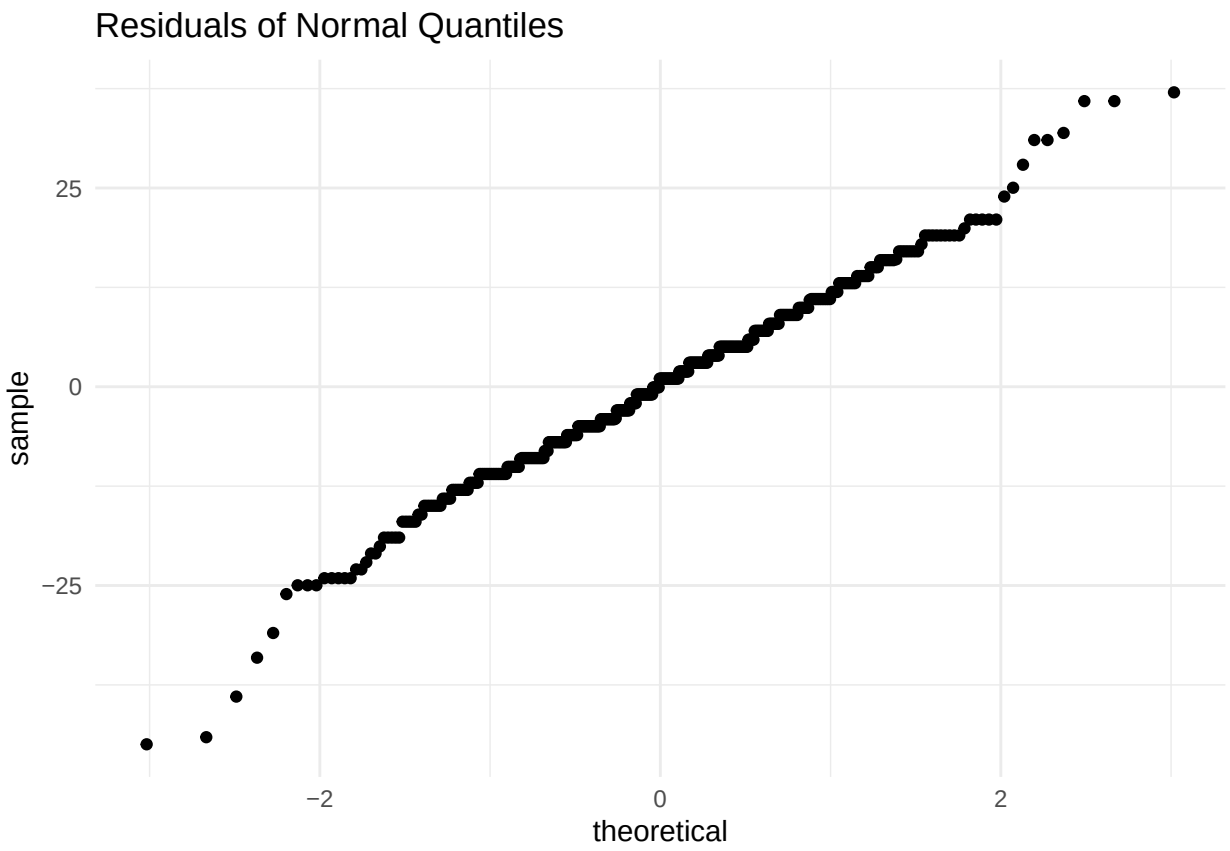
```
## # A tibble: 392 x 9
##    pregnant glucose pressure triceps insulin  mass pedigree   age diabetes
##       <dbl>   <dbl>    <dbl>   <dbl>   <dbl> <dbl>    <dbl> <dbl> <fct>
##  1        1      89       66      23      94  28.1    0.167    21 neg
##  2        0     137       40      35     168  43.1    2.29     33 pos
##  3        3      78       50      32      88  31      0.248    26 pos
##  4        2     197       70      45     543  30.5    0.158    53 pos
```

```
##  5        1     189       60       23     846  30.1    0.398    59 pos
##  6        5     166       72       19     175  25.8    0.587    51 pos
##  7        0     118       84       47     230  45.8    0.551    31 pos
##  8        1     103       30       38      83  43.3    0.183    33 neg
##  9        1     115       70       30      96  34.6    0.529    32 pos
## 10        3     126       88       41     235  39.3    0.704    27 neg
## # ... with 382 more rows
```

```
model_1 <- lm(pressure ~ diabetes, data=model_data)
model_data %>%
add_residuals(model_1) %>%
ggplot(aes(x=diabetes, y=resid)) +
geom_jitter() +
geom_boxplot() +
labs(y="residuals",
title="Residuals of Diabetes") +
theme_minimal()
```




```
model_data %>%
add_residuals(model_1) %>%
ggplot(aes(sample=resid)) +
geom_qq() +
labs(title="Residuals of Normal Quantiles") +
theme_minimal()
```

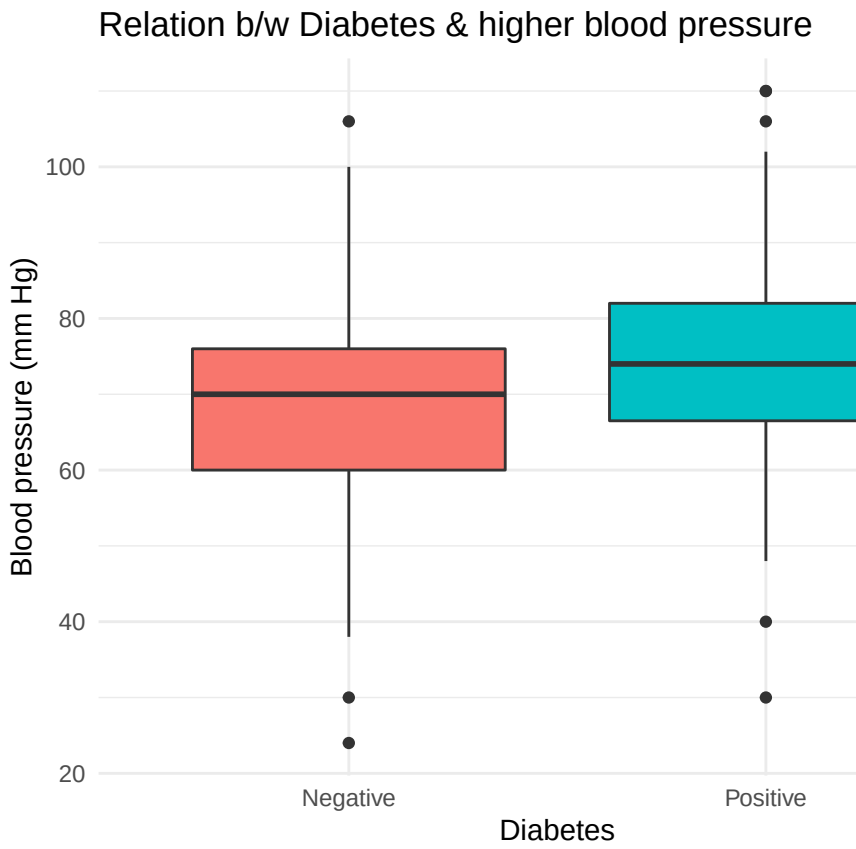



```
summary(model_1)
```

```
##
## Call:
## lm(formula = pressure ~ diabetes, data = model_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.969  -8.077   1.031   7.923  37.031
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  68.9695     0.7585  90.927  < 2e-16 ***
## diabetespos   5.1075     1.3172   3.878 0.000124 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.28 on 390 degrees of freedom
## Multiple R-squared:  0.03712,    Adjusted R-squared:  0.03465
## F-statistic: 15.04 on 1 and 390 DF,  p-value: 0.0001237
```

```
model_data %>%
mutate(Diabetes=recode(diabetes,
pos="Positive", neg="Negative")) %>%
ggplot(aes(x=Diabetes, y=pressure, fill=Diabetes)) +
```

```
geom_boxplot() +
labs(y="Blood pressure (mm Hg)",
title="Relation b/w Diabetes & higher blood pressure") +
theme_minimal()
```

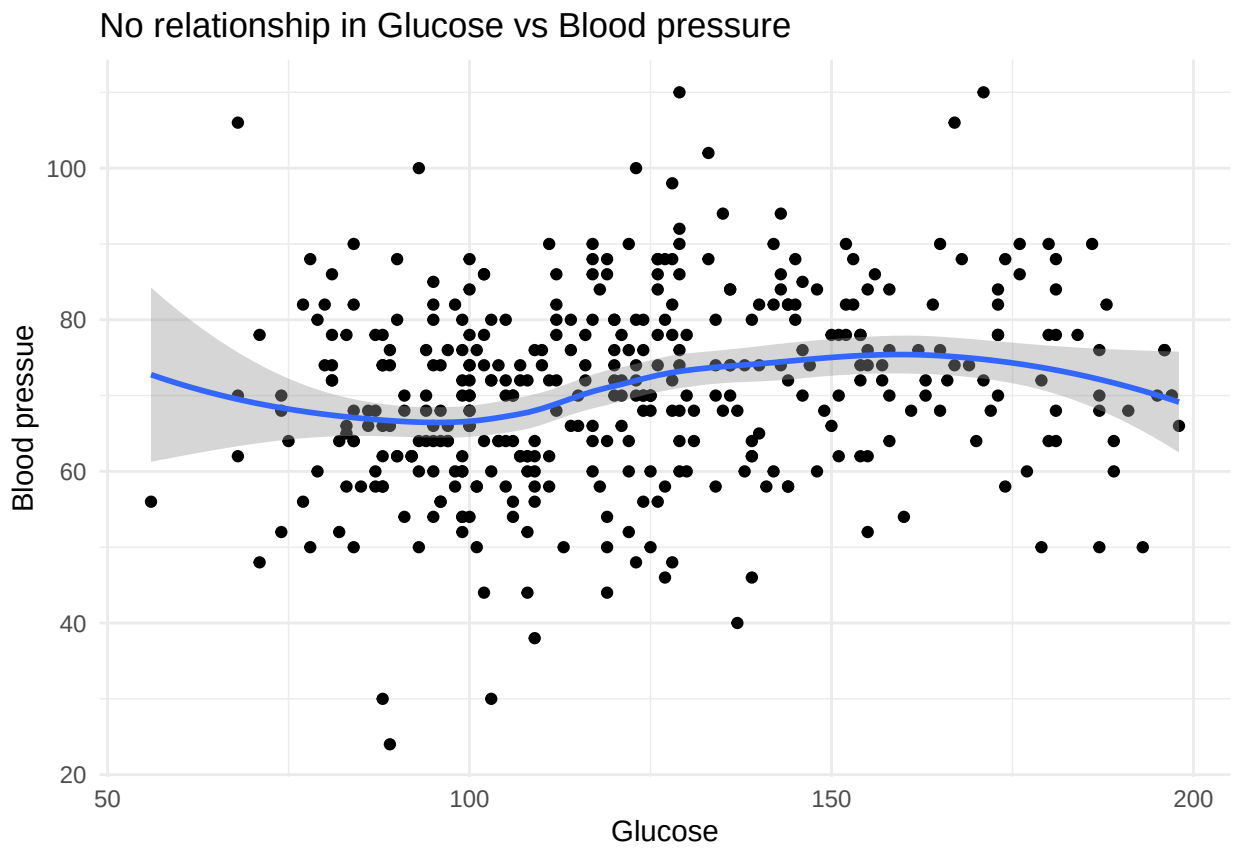


No violation of assumptions were observed.
Based on Null Hypothesis there is no observed difference in blood pressure between people with and without diabetes. Based on Alternate Hypothesis there is observed difference between blood pressure in people with and without diabetes. p=0.000124 (significance cutoff of 0.05), we don't want to consider null hypothesis. Higher blood pressure can be observed for people with diabetes.

Problem 4: We would like to consider 'glucose', 'insulin', 'triceps', 'mass', and 'age' as possible covariates. Plot them each against 'pressure' for consideration in the model as explanatory variables. Which variables would you consider including? Use AIC to select the best model that also includes 'diabetes' as a factor. Show your steps and reasoning, and then state the final model. Hint: AIC can be calculated using 'AIC()' or 'extractAIC()'; note that these two functions use different additive constants when calculating the likelihood, and so give differing values for AIC. However, they should lead to the same conclusions. You may find the 'step()' function useful as well.

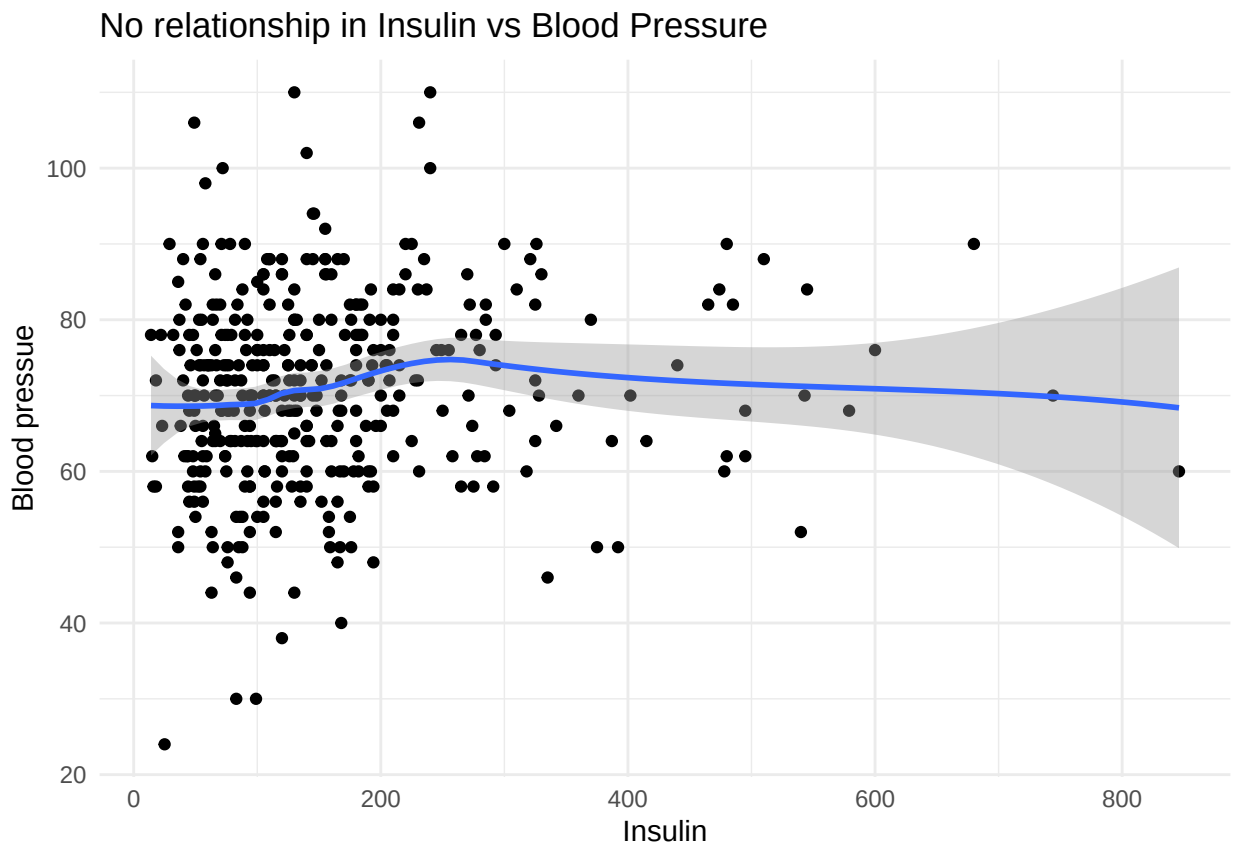
```
ggplot(model_data, aes(x=glucose, y=pressure)) +
geom_point() +
geom_smooth() +
labs(x="Glucose", y="Blood pressue",
title="No relationship in Glucose vs Blood pressure") +
theme_minimal()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



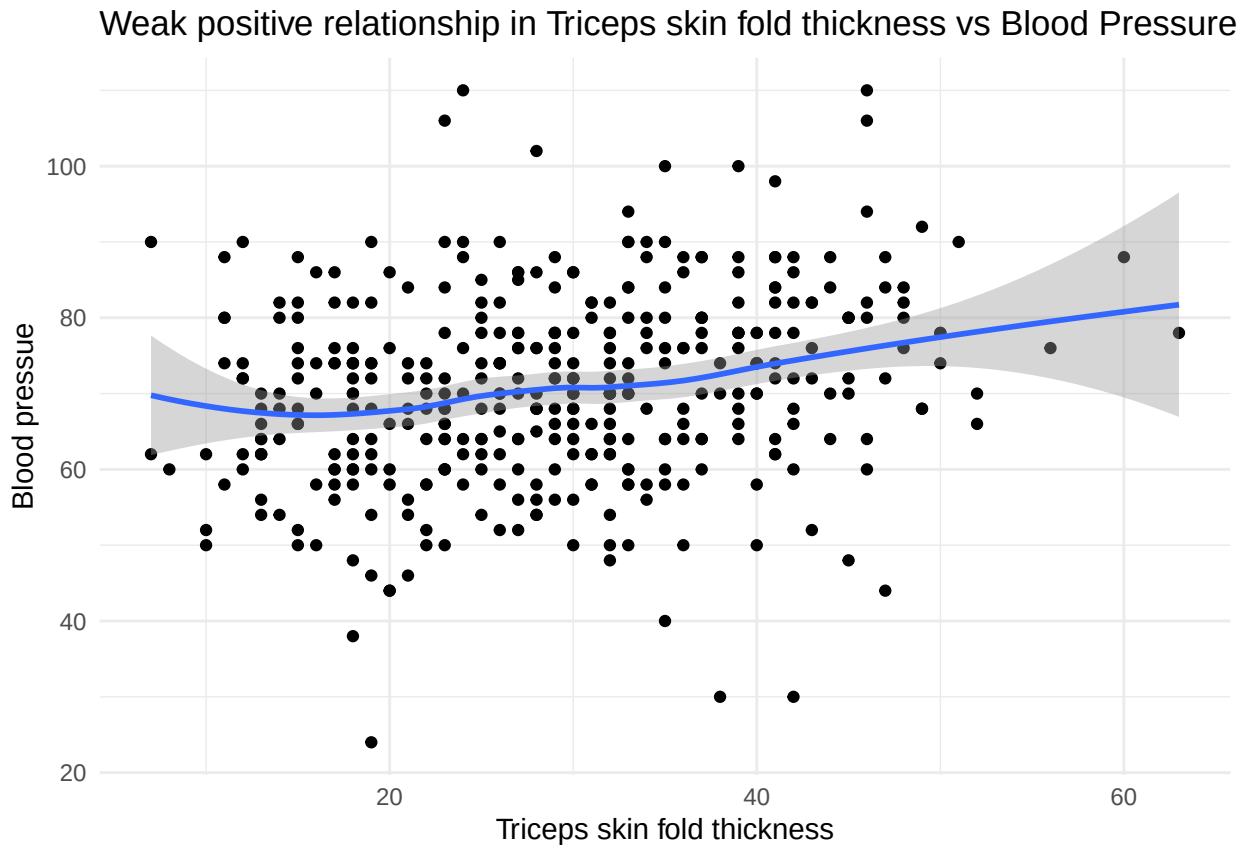

```
ggplot(model_data, aes(x=insulin, y=pressure)) +
geom_point() +
geom_smooth() +
labs(x="Insulin", y="Blood pressue",
title="No relationship in Insulin vs Blood Pressure") +
theme_minimal()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



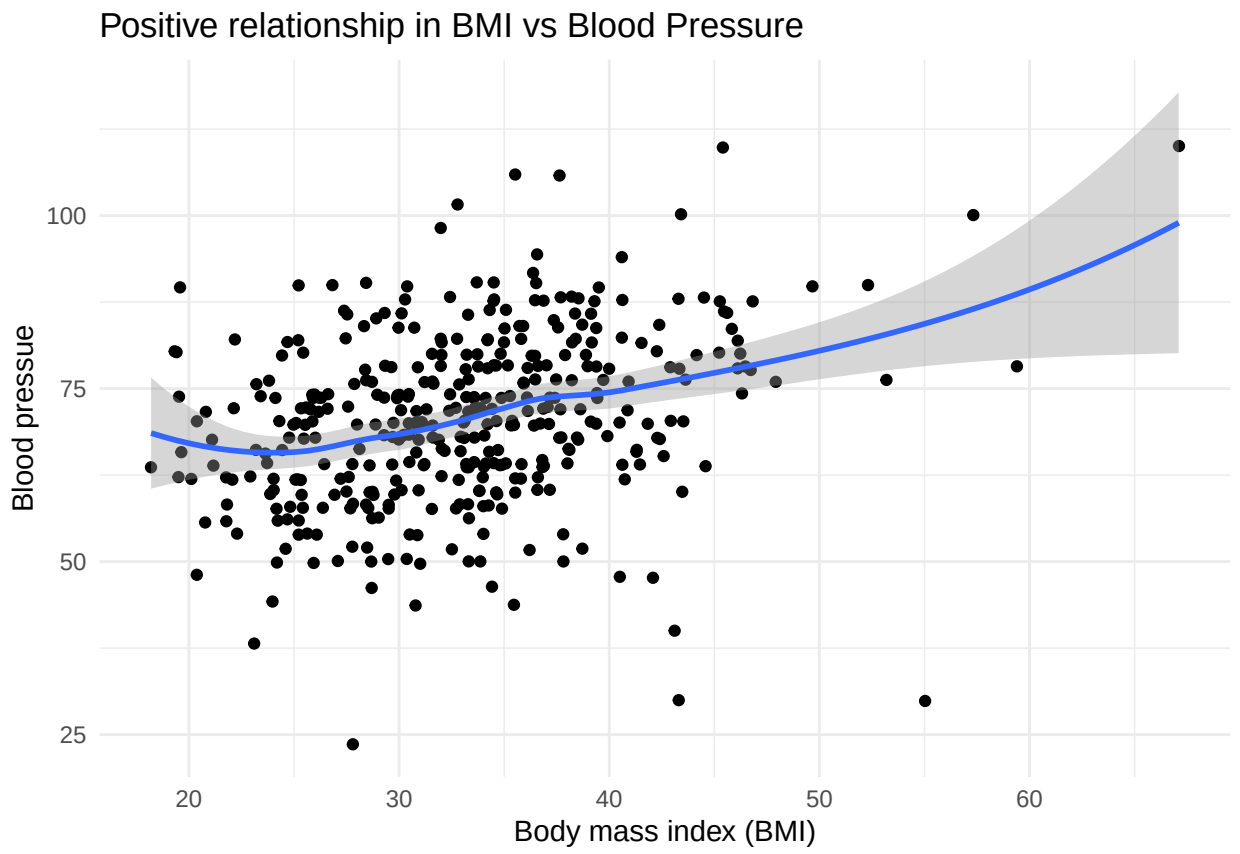

```
ggplot(model_data, aes(x=triceps, y=pressure)) +
geom_point() +
geom_smooth() +
labs(x="Triceps skin fold thickness", y="Blood pressue",
title="Weak positive relationship in Triceps skin fold thickness vs Blood Pressure") +
theme_minimal()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



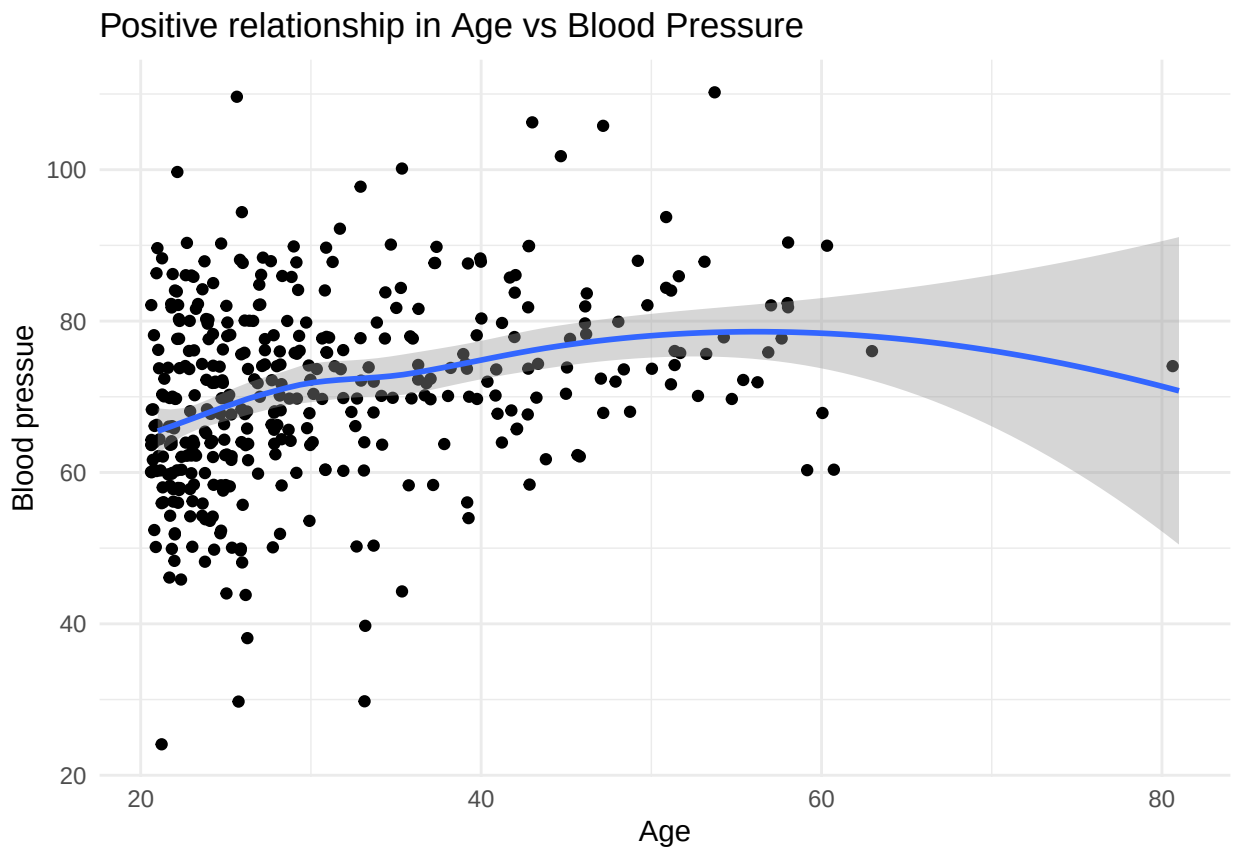

```
ggplot(model_data, aes(x=mass, y=pressure)) +
geom_jitter() +
geom_smooth() +
labs(x="Body mass index (BMI)", y="Blood pressue",
title="Positive relationship in BMI vs Blood Pressure") +
theme_minimal()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
ggplot(model_data, aes(x=age, y=pressure)) +
geom_jitter() +
geom_smooth() +
labs(x="Age", y="Blood pressue",
title="Positive relationship in Age vs Blood Pressure") +
theme_minimal()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Graph signifies that triceps have weak positive relationship, while age, bmi has positive relationship. We can consider Triceps as well if needed.

```
model_2 <- lm(pressure ~ diabetes + age + mass + triceps, data=model_data)
step(model_2)
```

```
## Start:  AIC=1915.33
## pressure ~ diabetes + age + mass + triceps
##
##            Df Sum of Sq   RSS    AIC
## - triceps   1       2.2 50607 1913.3
## - diabetes  1      23.5 50628 1913.5
## <none>                  50604 1915.3
## - mass      1    2682.6 53287 1933.6
## - age       1    3940.4 54545 1942.7
##
## Step:  AIC=1913.35
## pressure ~ diabetes + age + mass
##
##            Df Sum of Sq   RSS    AIC
## - diabetes  1      22.8 50629 1911.5
## <none>                  50607 1913.3
## - age       1    3988.0 54595 1941.1
## - mass      1    4420.4 55027 1944.2
##
## Step:  AIC=1911.52
## pressure ~ age + mass
```

```
##
##        Df Sum of Sq   RSS    AIC
## <none>              50629 1911.5
## - age   1    4768.6 55398 1944.8
## - mass  1    4929.7 55559 1945.9
```

```
##
## Call:
## lm(formula = pressure ~ age + mass, data = model_data)
##
## Coefficients:
## (Intercept)          age         mass
##     43.3129       0.3432       0.5065
```

```
model_3 <- lm(pressure ~ diabetes + age + mass, data=model_data)
AIC(model_1)
```

```
## [1] 3082.543
```

```
AIC(model_2)
```

```
## [1] 3029.777
```

```
AIC(model_3)
```

```
## [1] 3027.794
```

```
model_3
```

```
##
## Call:
## lm(formula = pressure ~ diabetes + age + mass, data = model_data)
##
## Coefficients:
## (Intercept)  diabetespos          age         mass
##     43.7050       0.5668       0.3345       0.4971
```

We will only consider age, mass as factors as adding AIC is not providing progressive results.

```
summary(model_3)
```

Problem 5: Use your model from Problem 4 to test the same hypotheses as Problem 3. State the null and alternative hypotheses, choose an alpha value, and state the p-value and your conclusions.Are your results the same or different? How do you explain this?

```
##
## Call:
## lm(formula = pressure ~ diabetes + age + mass, data = model_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.309  -7.193  -0.611   7.713  28.928
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.70503    3.32347  13.150  < 2e-16 ***
## diabetespos  0.56675    1.35607   0.418    0.676
```

```
## age          0.33445    0.06049   5.530 5.91e-08 ***
## mass         0.49711    0.08539   5.822 1.22e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.42 on 388 degrees of freedom
## Multiple R-squared:  0.1711, Adjusted R-squared:  0.1647
## F-statistic:  26.7 on 3 and 388 DF,  p-value: 1.011e-15
```

Based on Null hypothesis there is no observed difference in blood pressure (for both people with diabetes and without diabetes). However difference can be observed in blodd pressure between people with diabetes and without diabetes if based on alternative hypothesis. As the Significance cutoff of is 0.05, we will consider null hypothesis. Conclusion can be made that there is no relation between people with diabetes and without diabetes. Result varies from earlier as we considered BMI and age. The relationship doesnt exist if the age and BMI are same.